RSAC 2025 Conference

Many Voices. One Community.

SESSION ID: SAT-M05

Harnessing AI to Enhance Cloud Security While Addressing New Attack Vectors

George Gerchow

CSO Bedrock Security
https://www.linkedin.com/in/georgegerchow/

Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference LLC does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

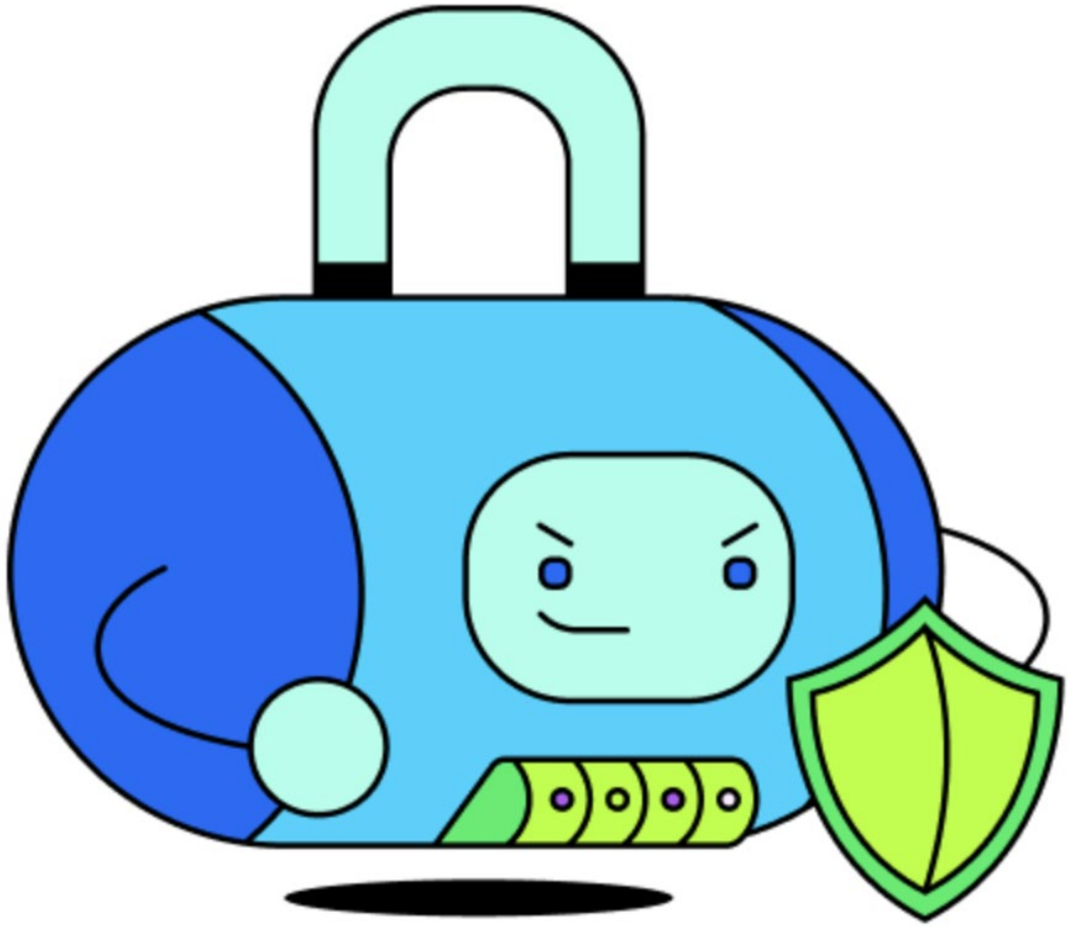
Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

© 2025 RSA Conference LLC or its affiliates. The RSAC and RSAC CONFERENCE logos and other trademarks are proprietary. All rights reserved.

If applicable, insert your organization's disclaimer statement here, in black (or delete this text block)

Guardian Bot (GB)

- We will cover this AI application later in the presentation
- It was created as an internal tool at MongoDB to automate GRC responses.
- The team had to fend off a series of possible AI attacks while building GB throughout its entire lifecycle



AI-powered bots, like **Guardian BOT (GB)**, are rapidly becoming essential applications in modern cloud security. These bots aren't just simple tools—they're dynamic systems that learn, adapt, and respond in real-time to emerging threats.

By leveraging AI, they offer proactive defense mechanisms, automate routine tasks, and help organizations scale their security operations. Today, we'll explore how AI bots, when integrated effectively, can transform cybersecurity by providing faster, more accurate responses while continuously evolving to address new challenges.

Introduction

AI is revolutionizing cybersecurity by transforming both offensive and defensive strategies.

This session will explore how AI has evolved to counter increasingly sophisticated, AI-driven attacks, focusing on practical examples from MongoDB's use cases.

Key Outcome: Learn how to apply AI-powered defenses in your organization to enhance security resilience.

The Evolution of AI in Cybersecurity

AI's use in threat detection has been around for years, but its capabilities are expanding at a rapid pace.

Modern AI systems can analyze vast behavioral data far beyond human capacity, dynamically adapting to new threats.

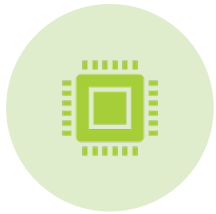
How You Can Apply It: Learn how AI can adapt in real-time to counter new attack vectors—apply these insights to keep your organization ahead of adversaries.

AI-Powered Threats

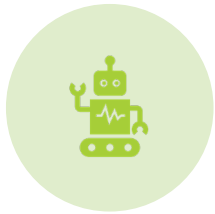
DEEPFAKES: AI-GENERATED MEDIA USED FOR IMPERSONATION.

PROMPT INJECTIONS: MANIPULATING AI SYSTEMS TO BEHAVE UNPREDICTABLY

AI-GENERATED PHISHING: HIGHLY SOPHISTICATED AND DIFFICULT-TO-DETECT PHISHING ATTEMPTS.



DATA POISONING: MALICIOUSLY CORRUPTING TRAINING DATA TO MISLEAD AI MODELS INTO MAKING FAULTY PREDICTIONS OR EXPOSING SENSITIVE DATA.



EDUCATE: UNDERSTAND THESE EVOLVING AI-DRIVEN THREATS.

LEARN: HOW TO IDENTIFY THESE RISKS IN YOUR ENVIRONMENT.

APPLY: IMPLEMENT STRATEGIES TO DEFEND AGAINST THESE THREATS USING AI-POWERED SOLUTIONS.

Real-Time AI Defenses

- **AI-driven Threat Intelligence**: Continuous monitoring and adaptation to evolving threats.
- **Behavioral Analytics**: Detecting anomalies and deviations in real-time, allowing for rapid response.
- **Automated Response Mechanisms**: AI-powered mitigation strategies to stop attacks before they escalate.
- **How You Can Apply It**: Implement real-time anomaly detection and automated defenses in your cloud environment to stay ahead of evolving attack strategies.

The Challenge

Growing Complexity in Security & Compliance

- Organizations face **an overwhelming volume of security & compliance inquiries** from customers, auditors, and regulators.
- **Manual processes** for responding to security questionnaires, RFPs, and compliance checks **are slow, inconsistent, and resource-intensive**.
- Cyber threats, including **AI-generated phishing, prompt injections, and data poisoning**, are **evolving too fast for traditional security measures** to keep up.

Why We Built Guardian Bot

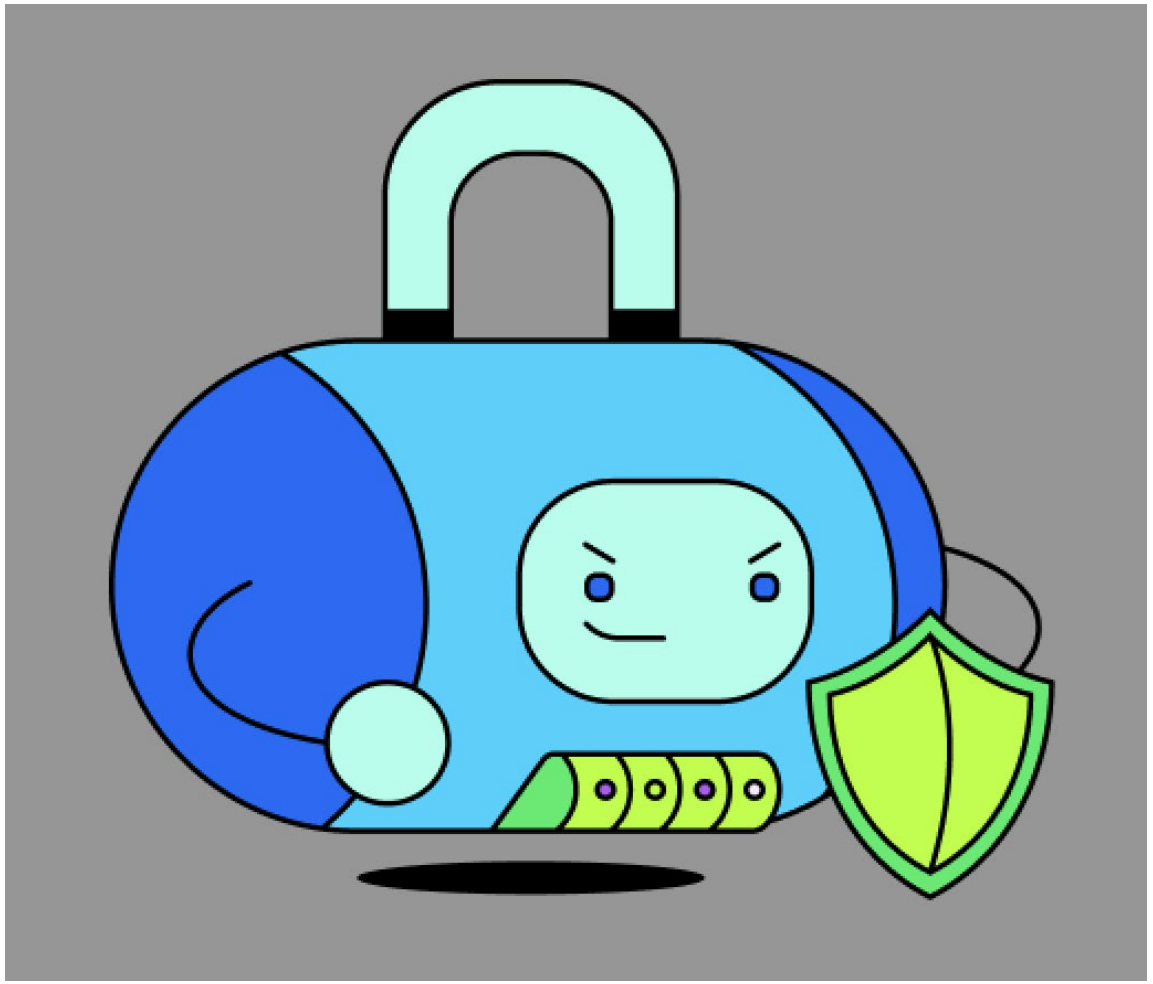
The Need for an AI-Driven Solution

Delays in security responses impact sales cycles, customer trust, and compliance timelines.

Security teams are **burdened with repetitive questions**, taking focus away from **critical risk mitigation efforts**.

Attackers are **leveraging AI to automate cyber threats**, requiring **real-time, AI-driven defenses** to counteract evolving attack vectors.

Introducing Guardian Bot: The Solution



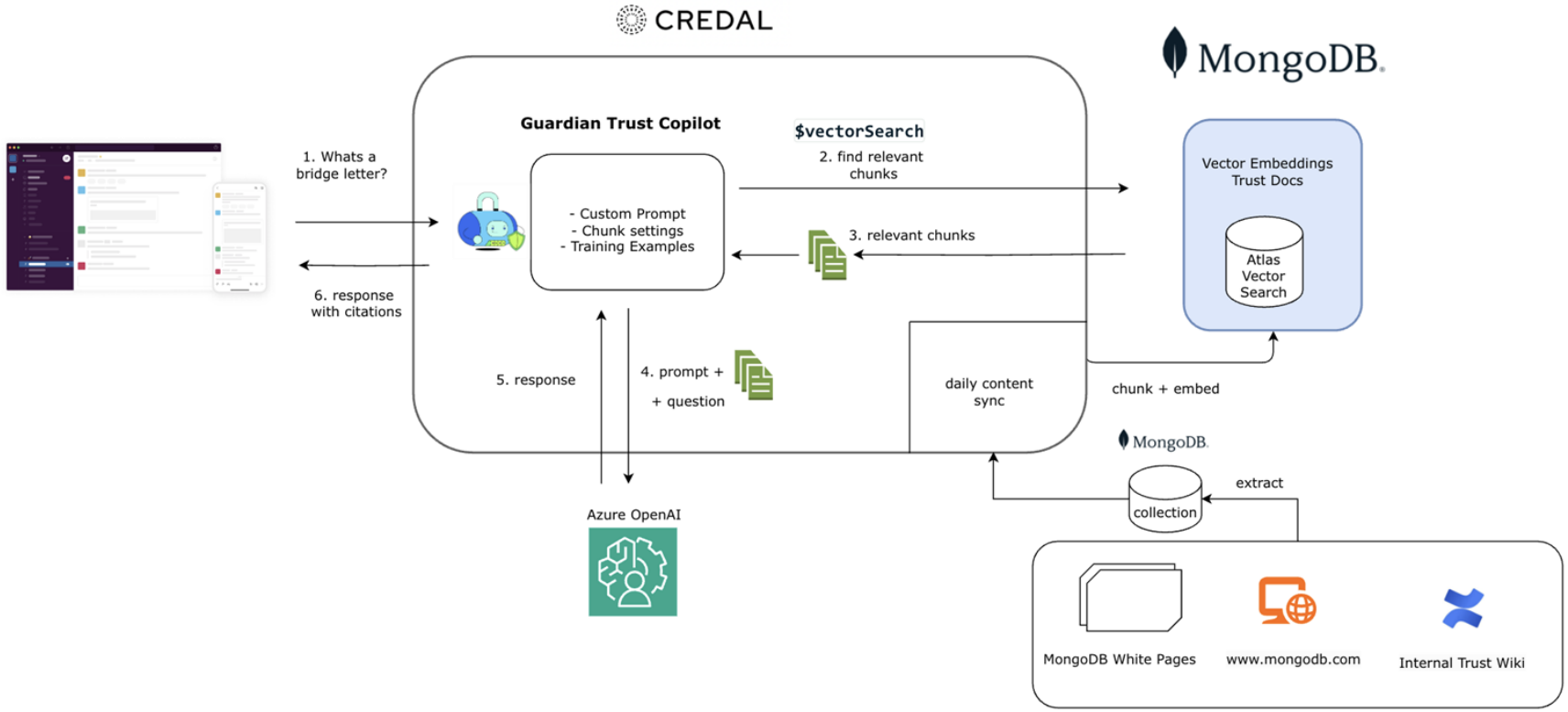
- **AI-powered automation** to provide **real-time, accurate security & compliance responses**.
- **Reduces response times from hours/days to seconds**, improving efficiency and customer satisfaction.
- **Enhances security resilience** by detecting and mitigating **AI-driven threats like data poisoning & adversarial attacks**.
- **Empowers security teams** to **focus on high-value tasks** by automating repetitive security and compliance workflows.

MongoDB Use Case – Guardian AI-Powered Security Response Bot

- *Guardian* is a next-generation Slack bot that enhances security response by streamlining queries related to security and compliance.
- Built on **Credal's LLM technology** and **Azure OpenAI GPT-4o**, Guardian accelerates response times while ensuring high-quality information delivery.
- **Outcome**: AI handles 87.5% of inquiries and 75% of follow-ups, freeing up human resources for strategic tasks.
- **Cost Savings**: $114,816 per year and 1,128 hours of time saved.
- **Efficiency Gains**: Improved speed, accuracy, and consistency in security responses.
- **Actionable Takeaway**: Implement AI-driven tools like Guardian to streamline your security operations and reduce operational costs.

How Guardian Bot Works

*The **Guardian Bot**, built on MongoDB's **central RAG platform** and supported by **Credal**, enables MongoDB's GTM teams to quickly access relevant, accurate information about security and compliance in under a minute.*



How Guardian Enhances Security

- **Improved Response Times**: Speeds up customer security inquiries for GTM teams.
- **Freed-up Security Personnel**: Allows security experts to focus on higher-priority tasks.
- **Cost Reduction**: Automates routine queries, saving operational costs.
- **How You Can Apply It**: Automate repetitive tasks in your own organization's security processes to increase efficiency and free up resources for strategic efforts.

Guardian in action

George Gerchow 9:39 AM
Does MongoDB have plans for FedRAMP High?

1 reply 3 days ago

Guardian APP Friday at 9:39 AM
Hey! Great question about FedRAMP High. Currently, MongoDB Atlas for Government has achieved FedRAMP Moderate Authorization. For any plans or updates regarding FedRAMP High, I would recommend reaching out to your Regional VP to discuss if there's a business case or more detailed information available.

If you need more details about our current FedRAMP status, you can check out our Trust Center here: FedRAMP – MongoDB Atlas For Government.

Hope this helps! Anything else you need?

George Gerchow Just now
Hello, does MongoDB have a bug bounty program?

1 reply

Guardian APP Just now
Hey! Yes, MongoDB does have a bug bounty program. We encourage security researchers and the community to report any security vulnerabilities they find in our products. You can find more details and participate in our bug bounty program by visiting our Trust Center: Trust Center.

Hope this helps! Anyt...
See more

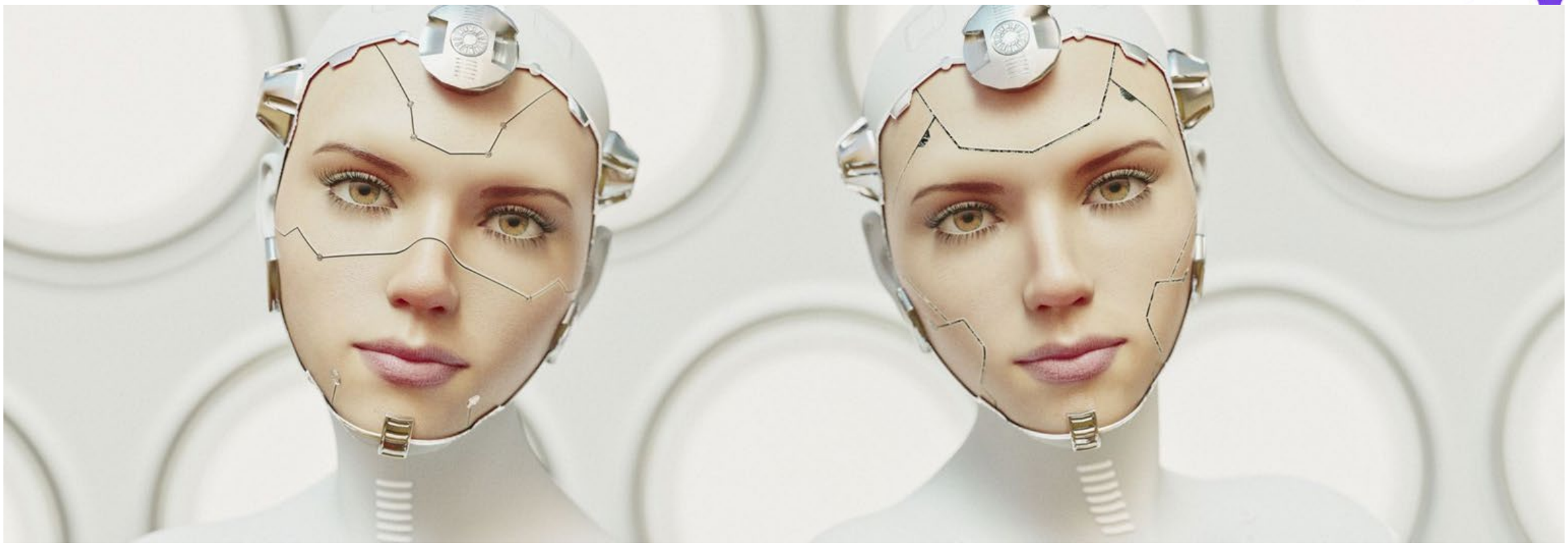
Guardian

👍 👎 ❌

Provide feedback or remove this message

MongoDB
MongoDB Atlas Trust Center
MongoDB is dedicated to securing and protecting your data through state-of-the-art technical and organizational security controls, numerous regulatory and compliance resources, and a growing collection of third-party attestations and certifications.

Using AI to Detect and Defend Against AI-Powered Attacks



- PROACTIVE DETECTION: AI IDENTIFIES ANOMALIES LIKE PROMPT INJECTIONS OR DATA POISONING ATTEMPTS IN REAL-TIME.
- CASE STUDY: LEARN HOW MONGODB'S GUARDIAN BOT HELPS IDENTIFY UNUSUAL PATTERNS AND DETECTS ADVERSARIAL THREATS THROUGH ANOMALY DETECTION MODELS.
- HOW YOU CAN APPLY IT: DEPLOY SIMILAR PROACTIVE AI MODELS TO DETECT AND NEUTRALIZE AI-DRIVEN ATTACKS IN YOUR ORGANIZATION.

Securing Guardian Bot

- ✅ **Hardened AI Training Data** – Implemented **data sanitization & adversarial training** to resist poisoning.
  ✅ **Real-Time AI Input Validation** – Integrated **strict prompt filtering** and **context-aware security controls**.
  ✅ **Differential Privacy Protections** – Prevented **model inversion attacks** by **limiting AI's ability to memorize sensitive data**.
  ✅ **Continuous Red Teaming & Bug Bounty Expansion** – Regular **adversarial testing and ethical hacking challenges** to keep the AI resilient.

Uncovering AI Threats

(Via Bug Bounties and Red Teaming)

📌 Phase 1: Prompt Injection Attacks

- 🎯 **What Happened?** Attackers attempted manipulated AI prompts to extract sensitive data.
- 🚨 **Example:** Queries attempted to **trick Guardian Bot into leaking security controls**.
- 🔒 **Fix:** Deployed **strict input validation & real-time context filtering**.

AI-powered adversarial inputs lead to poisoned training data

📌 Phase 2: Data Poisoning Attempts

- 🎯 **What Happened?** Malicious data was injected to corrupt AI decision-making.
- 🚨 **Example:** AI-generated compliance reports included **false security assertions**.
- 🔒 **Fix:** Introduced **data sanitization, adversarial training & anomaly detection**.

Corrupted AI models create risks of reverse engineering

📌 Phase 3: Model Inversion

- 🎯 **What Happened?** Attackers tried to reconstruct AI training data.
- 🚨 **Example:** AI responses were probed to **extract past compliance Q&As**.
- 🔒 **Fix:** Applied **differential privacy techniques & strict memory retention limits**.

Evasive AI-based threats attempt to bypass detection

📌 Phase 4: Adversarial AI Evasion Techniques

- 🎯 **What Happened?** AI-generated threats dynamically changed patterns to evade detection.
- 🚨 **Example:** Phishing messages were **altered in real-time** to avoid Guardian Bot's monitoring.
- 🔒 **Fix:** Enhanced **real-time anomaly detection & AI-driven behavioral monitoring**.

Continuous AI Threat Cycle Recap

1️⃣ Prompt Injection

➡️ Attack: Manipulate AI inputs.
➡️ Fix: Strict input validation.

💡 *Takeaway: Never trust user input—always validate.*

2️⃣ Data Poisoning

➡️ Attack: Inject bad training data.
➡️ Fix: Data sanitization & anomaly detection.

💡 *Takeaway: Guard your AI's learning process.*

3️⃣ Model Inversion

➡️ Attack: Extract training data.
➡️ Fix: Privacy controls & memory limits.

💡 *Takeaway: Protect sensitive data from being reconstructed.*

4️⃣ Adversarial Evasion

➡️ Attack: Bypass AI defenses.
➡️ Fix: Real-time anomaly detection.

💡 *Takeaway: AI threats evolve—your defenses must too.*

🔄 **Repeat & Adapt: AI security is never static**

Future of AI in Cybersecurity

- **AI-Driven Automation**: AI will continue to automate threat detection, response, and mitigation strategies..
- **Predictive Defense**: AI will increasingly predict and prevent attacks before they occur.
- **How You Can Apply It**: Future-proof your security strategies by incorporating AI to proactively detect, respond to, and mitigate emerging threats.

AI-Powered Data Protection

- **Automated Security** – AI classifies data, controls access and detects anomalies.
- **Adaptive Threat Defense** – Identifies and blocks threats like **data poisoning & insider attacks** in real time.
- **Privacy-Preserving AI** – Ensures compliance while securing data through encryption & federated learning.
- **Smarter Compliance** – AI simplifies security controls and automates risk assessments.

Key Takeaways

- **AI is essential** for modern cloud security in today's rapidly evolving threat landscape.
- **Threats are becoming more sophisticated,** requiring AI-driven defenses.
- **Zero Trust and CARTA models** enhanced by AI can provide stronger resilience against adversarial threats.
  - **Zero Trust** focuses on enforcing strict verification for all access requests.
  - **CARTA** builds on Zero Trust but adds **continuous monitoring and dynamic trust adjustments**.
- **Human Oversight Remains Critical**: Ensure proper human oversight when deploying AI in security operations to ensure a balanced, effective approach.

Apply Now

Immediate Actions (Post-Session)

- **Assess AI Security Readiness** – Review your organization's AI-driven security tools to identify potential gaps in **data poisoning defenses, anomaly detection, and automated response capabilities**.
- **Collaborate with Key Stakeholders** – Engage with **GRC, security engineering, and AI teams** to align on AI security risks and mitigation strategies.

Next 3 Months Implementation Plan

1

Implement AI-Based Anomaly Detection – Deploy **real-time AI monitoring** for **prompt injections, adversarial AI threats, and unauthorized behavioral deviations**.

2

Harden AI Training Data Against Poisoning – Introduce **data sanitization techniques, adversarial training, and differential privacy methods** in your AI models.

3

Automate Compliance & Security Responses – Leverage AI-driven bots (like Guardian Bot) to reduce manual workloads in **customer security inquiries, compliance reporting, and incident response**.

4

Evaluate Zero Trust & CARTA AI Enhancements – Conduct an internal **Zero Trust & CARTA gap assessment** to identify where AI-driven policy enforcement and **continuous risk assessment** can improve defenses.

5

Conduct AI Security Simulations & Bug Bounty Testing – Test AI models for **vulnerabilities (e.g., data poisoning, model inversion)** using **red teaming and bug bounty programs**.

RSAC 2025 Conference

Many Voices.
One Community.

Thank You